Notes -- Chapter 4

Relations in categorical data

To analyze categorical data we will use counts or percents of individuals that fall into various categories.

This is a TWO-WAY table because it describes two categorical variables. Education is the ROW variable and age is the COLUMN variable. The entries in the table are counts of persons in each group. The row totals are given at the right and the column totals are given at the bottom of the table.

Table 4.6 Years of school completed by age (thousands of persons) pg. 216

	Age group			
Education	25 to 34	35 to 54	55 and over	Total
Did not complete high school	5325	9152	16035	30512
Completed high school	14061	24070	18320	56451
1 to 3 years of college	11659	19926	9662	41247
4 or more years of college	10342	19878	8005	38225
Total	41388	73028	52022	166438

Notes -- Chapter 4

Interpreting Correlation and Regression

Some cautions:

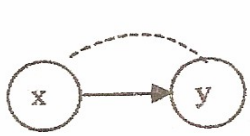
- Correlation and regression describe only LINEAR RELATIONSHIPS.
- Both r and the least squares regression line can be strongly influenced by a few extreme observations.
- ALWAYS plot your data before interpreting regression or correlation.
- Don't make predictions based on a linear regression for values of x that are outside the domain of the explanatory variable. (Extrapolation) Such predictions cannot be trusted.
- The correlation and regression we have studied is for two variables at a time. Sometimes there are other variables (that we did not measure or even consider) that have an important effect on the relationship between the variables in the study. Such a "lurking" variable can falsely suggest a strong relationship between variables or even hide a relationship that is really there.
- If a study uses averaged data rather than data on individuals don't use the results of the study to draw conclusions about individuals. Data from individuals usually shows more scattter from the regression line than averaged data and, therefore, a lower correlation. CORRELATIONS BASED ON AVERAGES ARE USUALLY TOO HIGH WHEN APPLIED TO INDIVIDUALS.

- A strong association between variables is not enough to draw conclusions about cause and effect. A strong association may reflect any of these three relationships.

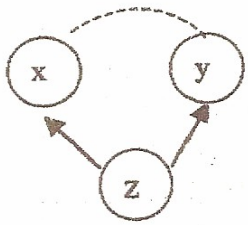
    - Causation: changes in x cause changes in y . If we can change x, then we can cause a change in y.

    - Common response: both x and y respond to changes in some unobserved variable(s). Sometimes we can predict changes in y from x, but changing x does not cause a change in y.

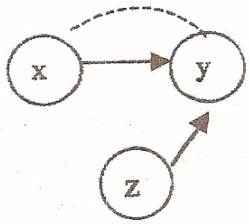
    - Confounding: the effect of x on y is mixed up with the effects on y of other varaibles.



Causation



Common response



Confounding

ASSOCIATION DOES NOT IMPLY CAUSATION.

The best way to get evidence of causation is to do a controlled experiment in which x is changed and lurking variables are kept under control. We will observe changes in y that are actually a result of changes in x.